PROMPTING FOR ROBUSTNESS: EXTRACTING ROBUST CLASSIFIERS FROM FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

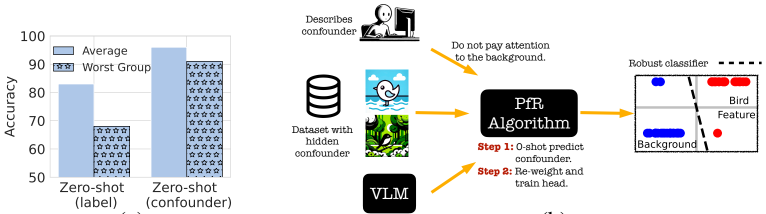
Machine learning models can fail when trained on distributions with hidden confounders (spuriously correlated with the label) and tested on distributions where such correlations are absent. While numerous algorithmic solutions have been explored for such distribution shifts, a surprisingly effective way to empirically improve robustness on some other types of shift (*e.g.*, Imagenet and its distribution shifts) is to use stronger open-vocabulary classifiers derived from foundation models. In this work, we note that for more controlled shifts regulated by spurious correlations, the zero-shot and few-shot performance of foundation models is no better than ERM models, and remains unchanged when pretrained data/model is scaled. However, even in those situations, they are quite accurate at predicting possible confounders. We leverage this observation to propose Prompting for Robustness (PfR) which first uses foundation models to zero-shot predict the confounder on given labeled examples, and then learns a classifier with balanced performance across different groups. In a simplified setup, we theoretically analyze the zero-shot behavior of multimodal models explaining how contrastive pretraining can learn features that strongly couple the confounder with more robust features. Across 5 vision and language tasks, we show that PfR's performance nearly equals that of an oracle algorithm (group DRO) that leverages labeled spurious attributes.

1 INTRODUCTION

Machine learning classifiers are often trained on datasets with hidden confounders spuriously correlated with the label. Since ERM models latch onto these confounders and fail catastrophically on underrepresented (minority) groups where the confounder is uncorrelated with the label, numerous algorithms have been proposed to make ERM more robust to such confounders (*e.g.*, Ben-Tal et al., 2013; Arjovsky et al., 2019; Liu et al., 2021). On the other hand, driven by the unprecedented zero-shot prediction capabilities of foundation models, the common strategy of learning classifiers has been to simply prompt them with class names directly Wei et al. (2020); Brown et al. (2020). In fact, zero-shot prompting sometimes yields classifiers that are more robust than ERM classifiers trained on downstream data (Hendrycks et al., 2020; Fang et al., 2022), *e.g.,* as seen in robustness gains observed on benchmarks like ImageNet with distribution shifts Radford et al. (2021). However, as we show in our work, such gains do not proportionately transfer to other forms of distribution shift such as when confounders that are highly predictive of the label in training distribution are no longer correlated with the label on test (Yang et al., 2023; Tu et al., 2020; Hall et al., 2023).

In this work, we aim to improve the performance of foundation models on data paritions (groups) where the confounder is not correlated with the label (minority group). One way is to incorporate downstream labeled data. Unfortunately, unless we have access to deconfounded data (without the spurious correlation), fine-tuning naïvely would result in the same issues as standard ERM training, as we confirm experimentally. However, with open-vocabulary foundation models, we can provide for robustness by *telling* the model about the confounder directly (*i.e.*, by describing it in the classification prompt). However, we observe that even this doesn't improve zero-shot robustness (see Sec. 2).

We make an intriguing observation: while foundation models are not robust zero-shot classifiers of the true label, they perform remarkably well in predicting the *presence* of spurious attributes. Moreover, we observe that while scaling up the model size and pretraining data does not improve the performance of label prediction on minority groups, the worst group performance of spurious attribute



(a) (b)

Figure 1: (a): *Foundation models are not robust to spurious correlations, but can predict them*; Averaged across 4 tasks with spurious correlations, we see that while foundation models perform much worse on groups where the spurious correlation is absent, they are highly accurate at predicting the spurious attribute itself, across all groups. (b): *Prompting for Robustness (PfR)*: Leveraging this we propose our method PfR that learns robust classifiers from foundation models in two steps. In Step 1, PfR prompts foundation models to zero-shot predict the spurious attribute on a labeled dataset with spurious correlations, and in Step 2 it learns a robust classifier by minimizing worst group loss, across groups given by the combination of the predicted attribute and label.

prediction does. Motivated by these findings, we propose a simple technique that we call *Prompting for Robustness (PfR)*. PfR learns robust classifiers for downstream tasks with a few labeled examples and a language description of the confounding attribute. PfR first uses the language description to prompt for a zero-shot classifier that accurately predicts the spurious feature on each labeled examples. The value of the label and the predicted confounder jointly define a set of disjoint groups in our data. Then, a robust predictor is learnt by minimizing worst group loss, similar to group DRO, as described by Sagawa et al. (2019), but without ground-truth knowledge of examples in the minority group. This simple method yields surprising performance gains of $\geqslant 40\%$ (averaged across datasets) relative to zero-shot performance of foundation mdoels and ERM on downstream data alone. We further illustrate the applicability of our findings by showcasing its efficacy in extracting group annotations for auditing zero-shot (or ERM) models to assess their robustness. Specifically, we prompt GPT-4V to annotate Chest-Xray 14 dataset (Wang et al., 2017) for the presence of chest drains (the spurious attribute) and observe a significant robustness gap among ERM models. Finally, in a simplified setup for multimodal contrastive pretraining, we theoretically show that when the spurious correlations in the downstream task are also present in the pretraining distribution over image, and text pairs, then contrastive pretraining learns: (i) image features that couple the spurious feature with other robust features, while placing a higher weight on the spurious one; and (ii) text features that are almost identical for the text descriptions of the label and the spurious attribute.

In summary our key contributions are as follows. First, we study the performance of foundation models across five vision and language classification tasks with hidden confounders, and observe that while foundation models have poor zero-shot performance on minority examples (that does not improve with scale), they are accurate at predicting the value of the confounder. We confirm these findings theoretically in a simplified setup for multimodal contrastive pretraining. Second, we leverage this finding to propose a simple method: PfR which first zero-shot predicts the confounder when given a text description of it, and then learns a robust classifier across predicted groups. Empirically, we show PfR's worst group performance nearly matches the oracle (group DRO) on all datasets.

Problem setup. For a classification task, we use $\mathcal{X}$ to denote input set of text/image and $\mathcal{Y}$ for the set of labels. We also define a set $\mathcal{C}$ for spurious attributes (also called confounders). With $\mathcal{G} =: \{G_1, G_2, \ldots, G_k\}$, we define a set of groups where each G_i corresponds to a unique pair of label and confounder values (y_i, c_i). Under distribution $P(x, y, c)$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{C}$, the average error of a label classifier f is $\mathrm{err}_{\mathrm{y}}^{\mathrm{av}}(f) =: \mathbb{E}_P\left[\mathbb{1}(f(x) \neq y)\right]$ and spurious atribute classifier g is $\mathrm{err}_{\mathrm{c}}^{\mathrm{av}}(g) =: \mathbb{E}_P\left[\mathbb{1}(g(x) \neq c)\right]$. Similarly, their corresponding worst-group errors are: $\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) =: \max_{G \in \mathcal{G}} \mathbb{E}_{P|G}\left[\mathbb{1}(f(x) \neq y)\right]$ and $\mathrm{err}_{\mathrm{c}}^{\mathrm{wg}}(g) =: \max_{G \in \mathcal{G}} \mathbb{E}_{P|G}\left[\mathbb{1}(g(x) \neq c)\right]$. We define the *robustness gap* for any predictor as the difference between the average and the worst group errors for it. In this work, our goal is to learn a label classifier with (i) high average accuracy, and (ii) low robustness gap. For this, we are given a text description t_c of the confounder c, and few *i.i.d.* samples $\mathcal{D}$ from $P(x, y)$. Unless specified, we assume that group annotations are not given to us. We use FM to denote a foundation model, whose prediction of the spurious attribute in x is $\mathrm{FM}(x, t_c)$.

2 Zero-shot robustness of foundation models

Large zero-shot performance gap between the average and worst group. Zero-shot results are in Table 2. When evaluating CLIP L/14 models on vision datasets, a notable drop of 32% is observed

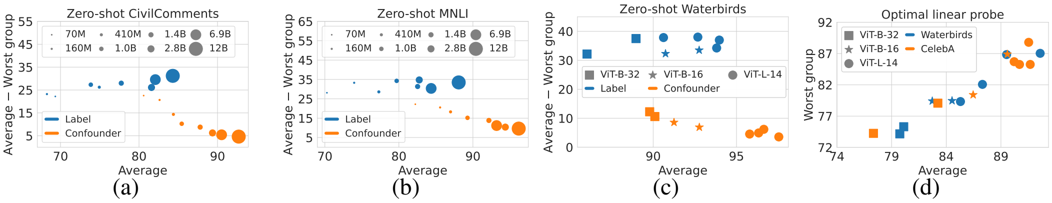


Figure 2: *Robustness gap versus average performance as pretraining data and model sizes increase.* We observe that while the robustness gap for confounder prediction decreases the gap between average and worst case increases or remains the same for label prediction.

between average and worst group accuracy on Waterbirds dataset, and a drop of 3.5% is observed on CelebA. Turning to language datasets, the evaluation of the Llama-2 13b model indicates a significant 25% performance decline in CivilComments and a 7% drop in MNLI. Notably, the drops observed here are similar to the performance drops observed with models trained with ERM on their corresponding labeled data (Sagawa et al., 2019; Idrissi et al., 2022). The decline seen with ERM models is typically ascribed to the existence of hidden confounders in the training data (Sagawa et al., 2019), suggesting that pretraining datasets also frequently suffer from analogous spurious correlations. We formalize this intuition in Sec. B.

Incorporating the group description naïvely does not help out of the box. We incorporate spurious attribute description in our zero-shot prompt to predict the label and the spurious attribute jointly. Results are shown in Table 1. However, the zero-shot performance for the worst-case group doesn't improve – there is less than a 1% change between the zero-shot and zero-shot with spurious attribute description rows in Table 1. We also evaluated other variants, where we explicitly instructed the model to ignore spurious attributes, but this did not impact worst-group performance (see App. G.2).

Foundation models are surprisingly good at predicting the presence of hidden confounders. Results are in Table 1. Instead of incorporating spurious attribute description together with the label, we experiment with predicting the presence of a spurious attribute alone. On all standard spurious correlation benchmarks, we observe that the average performance of predicting the presence of the spurious attribute is around 95% with a similar worst-case group performance. This consistent performance is observed across different groups, emphasizing that, despite foundation models exhibiting significant robustness gaps in the joint prediction of spurious attributes and labels, the predictive accuracy for spurious attributes alone remains superior.

Scaling pretraining datasets and models does not improve zero-shot group robustness. The scaling trend results are presented in Fig. 2 (a)-(c), showcasing the performance plotted on average against the difference between average performance and worst-case performance. We analyze this difference in comparison to the average case for both zero-shot label and spurious attribute prediction. As we scale up the pretraining datasets and models, we observe that while the difference reduces for the cofounder prediction, the difference doesn't improve for the label prediction task.

Scaling pretraining datasets and models does improve underlying representations. As expected we observe that the average and worst-case accuracy (trained with DRO on downstream labeled data) improves as we increase the scale of model size and pretraining data (Fig. 2 (d)).

CXR-Drain: Annotating confounders with GPTV-4. We evaluate the ability to predict spurious correlation in a zero-shot way on a task where ground truth annotations are not publicly available. We choose to annotate 2400 images from Chest Xray-14 dataset (Wang et al., 2017) for the presence of chest drain with GPT4-V (details in App. G.3). On this dataset, the goal is to predict the whether the patient suffers from pneumothorax disease and the tube in the chest cavity acts as a confounder.

Prompt	Predict	Waterbirds		CelebA		CivilComments		MNLI	
		WG	Avg	WG	Avg	WG	Avg	WG	Avg
Is this label L?	L	59.38	91.97	77.69	81.11	59.25	85.75	76.54	84.79
Is this label L? Ignore confounder C.	L	61.37	92.58	86.73	90.28	52.81	87.41	77.95	80.56
Is this label L and confounder C?	L,C	57.38	88.15	78.54	83.11	54.29	86.60	75.73	82.91
Is this confounder C?	C	90.55	96.33	95.01	99.15	86.73	92.70	92.37	96.19

Table 1: *Naïvely incorporating the confounder description into the label classification prompt does not improve robustness.* Results on four datasets with known spurious attributes.

We observe that models trained with ERM show a significant performance gap on the constructed CXR-Drain dataset (Table 2). Further, the worst group is not the group with least samples, and hence, re-weighting based methods (Idrissi et al., 2022; Kirichenko et al., 2022) would perform poorly. Due to its unique properties, we believe that CXR-drain will also serve as a crucial benchmark for future research on spurious correlations, and we plan to publicly release the dataset.

3 Prompting for Robustness

Our results in Section 2 suggest that zero-shot classification with foundation models often attains high average group accuracy but low worst-group accuracy. However, we note that they are surprisingly accurate at predicting the presence of a confounder. We leverage this finding to propose a simple but effective method: Prompting for Robustness (PfR). PfR learns a robust classifier given a few labeled examples and a text description of the confounder.

Method	**Waterbirds**		**CelebA**		**CivilComments**		**MNLI**		**CXR-Drain**	
	WG	Avg	WG	Avg	WG	Avg	WG	Avg	WG	Avg
Zero-shot	59.38	91.97	77.69	81.11	59.25	85.75	76.54	84.79	–	–
ERM	70.71	98.75	54.84	94.96	61.35	92.42	67.30	87.71	51.79	76.10
JTT	85.86	95.47	82.49	92.74	72.73	90.54	72.75	86.73	56.52	77.53
PfR (ours)	**91.05**	94.32	**88.05**	91.97	**77.83**	88.70	**81.28**	84.60	**68.55**	76.73
Group DRO (oracle)	93.23	94.40	90.79	92.32	80.21	86.52	81.54	84.37	–	–

Table 2: *PfR improves worst group performance over ERM and zero-shot foundation models:* On five benchmarks from Section 2 we evaluate average and worst-group performance of PfR and compare it with baselines JTT, ERM, and zero-shot.

Prompting for Robustness (PfR). PfR (summarized in Algorithm 1) runs in two stages. In the first stage, PfR prompts an open vocabulary foundation model FM with the text description t_c of the confounding attribute and recovers a zero-shot prediction of the confounder c on any given input (for *e.g.,* in the case of CivilComments the confounder is described as "race, religion or gender"). Using this, each training example (x_i), which was previously annotated only for the label of interest (y_i), is additionally annotated with the value of the confounding attribute $(\widehat{c}_i)$ (for *e.g.,* "black/white and christian/muslim"). The training dataset is then split into disjoint groups $\widehat{\mathcal{G}}$ based on the paired value $(y_i, \widehat{c}_i)$ of the label and predicted confounder. In the second stage, PfR learns a robust classifier by minimizing the worst group loss over each predicted group, minimizing

$$\min_{f} \max_{G \in \widehat{\mathcal{G}}} \; \mathbb{E}\left[\ell(f(x), y) \mid x \in G\right]. \tag{1}$$

The above objective can be optimized with an online algorithm that treats f and G as players in a minimax game, analogously to the group DRO algorithm described by Sagawa et al. (2020). Hence, we reuse their Algorithm 1 to optimize our objective in Equation (1).

Setup and baselines. On the language tasks we use Llama2-7b/13b models Touvron et al. (2023) for zero-shot prediction (reporting max of the two), and on the vision tasks we use CLIP-ViT-L/16 Radford et al. (2021). We compare to JTT Liu et al. (2021), a prior method for robustness that does not require group labels, as well as standard ERM. We also include Group DRO Sagawa et al. (2019) as an oracle baseline that has access to true group labels. All few-shot methods including PfR are used to train a linear head over fixed features. In the language task we train a linear head on top of features learned by finetuning a RoBERTa encoder Liu et al. (2019) on the MNLI/CivilComments dataset, and for vision tasks we train a linear head over CLIP's image encoder.

Results. In Table 2, we compare average and worst group performance for different methods. First, we observe that averaged across datasets, PfR reduced worst group error by 47% compared to zero-shot, and 52% and 30% compared to ERM and JTT, respectively. On some datasets like Waterbirds, the worst group gains are as high as $> 75\%$. More importantly, PfR's performance closely matches that of the oracle Group DRO algorithm across all datasets. Additionally, unlike overly pessimistic DRO objectives like CVaR-DRO Hu et al. (2018), the average performance is not significantly compromised from trying to improve worst group accuracy. Thus, we see that PfR learns a classifier robust to spurious correlations without much human annotation overhead beyond a description of the confounder.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yujia Bao and Regina Barzilay. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*, 2022.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Charles-Alban Deledalle, Loic Denis, Sonia Tabti, and Florence Tupin. *Closed-form expressions of the eigen decomposition of 2 x 2 and 3 x 3 Hermitian matrices*. PhD thesis, Université de Lyon, 2017.

John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406, 2021. doi: 10.1214/20-AOS2004. URL `https://doi.org/10.1214/20-AOS2004`.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Saurabh Garg, Amrith Setlur, Zachary Chase Lipton, Sivaraman Balakrishnan, Virginia Smith, and Aditi Raghunathan. Complementary benefits of contrastive learning and self-training under distribution shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19338–19347, 2023.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit disparities between gender groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2778–2785, 2023.

Jeff Z HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. *arXiv preprint arXiv:2211.14699*, 2022.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Jeff Z HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *arXiv preprint arXiv:2204.02683*, 2022.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34: 309–323, 2021.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*, 2022.

Yoonho Lee, Michelle Lam, Helena Vasconcelos, Michael Bernstein, and Chelsea Finn. Interactive model correction with natural language. In *XAI in Action: Past, Present, and Future Applications*, 2023.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Amrith Setlur, Don Dennis, Benjamin Eysenbach, Aditi Raghunathan, Chelsea Finn, Virginia Smith, and Sergey Levine. Bitrate-constrained dro: Beyond worst case robustness to unknown group shifts. *arXiv preprint arXiv:2302.02931*, 2023.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 19847–19878. PMLR, 2022.

Nimit Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Ré. Barack: Partially supervised group robustness with guarantees. *arXiv preprint arXiv:2201.00072*, 2021.

Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. *arXiv preprint arXiv:2304.03916*, 2023.

Runtian Zhai, Chen Dan, Arun Suggala, J Zico Kolter, and Pradeep Ravikumar. Boosted cvar classification. *Advances in Neural Information Processing Systems*, 34:21860–21871, 2021.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. *arXiv preprint arXiv:2306.04272*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Appendix

Appendix Outline

A. Additional details and observations from our zero-shot experiments in Sec. 2.

B. Theoretical analysis of multimodal contrastive pretraining.

C. Additional experiments and theoretical analysis of PfR

D. Related Work.

E. Future Work.

F. Proofs for our theoretical results.

G Details on zero-shot prompts.

A Additional details and observations from our zero-shot experiments in Sec. 2.

A.1 Setup

Datasets. We experiment with datasets in both language and vision modalities. For language, we experiment with: (i) MNLI (Williams et al., 2017), where the prediction task is relationship between two input sentences as being contradiction, entailment, or none of the two. Here the spurious attribute is the presence of negation words, e.g., 'no', and 'never'. (ii) CivilComments (Borkan et al., 2019; Koh et al., 2021), where the task is toxicity prediction and the spurious correlation is with the underlying attribute annotating the comment, e.g., male versus female, Christian versus Muslim, etc. For the vision modality, we experiment with: (iii) Waterbirds (Sagawa et al., 2019), where the prediction task is water bird versus land bird classification, and the spurious attribute is the background of the image (i.e., land versus water background); (iv) CelebA (Sagawa et al., 2019), where the prediction task is gender and the spurious attribute is the color of hair. We also experiment with the CXR-drain dataset introduced in Sec. 2.

Experimental setup. For our zero-shot probing results, we experiment with a number of pretrained foundation models. For vision, we experiment with CLIP (Radford et al., 2021; Gadre et al., 2023). For language, we experiment with RoBerta (Liu et al., 2019), Llama-2 (Touvron et al., 2023) and Pythia models (Biderman et al., 2023). We also experiment with publicly available models where we vary the model and pretraining dataset sizes in each category. For our ERM experiments, we train linear classifiers on the penultimate layer outputs (representation). For our zero-shot probes, we leverage standard prompts commonly used in the literature. Precise details about prompts used on each dataset are in App. G.

Evaluation metrics. Along with the prediction accuracy of the label on the worst-case group, we also report average performance. Additionally, we also evaluate the performance of predicting the spurious attribute.

B Theoretical analysis of multimodal contrastive pretraining

From Section 2, we recall that the worst group zero-shot performance in some cases (like predicting the label of a task with hidden confounders) never improves with scale. So, why does confounder prediction improve? In this section, we analyze both these trends theoretically when pretraining on data where the label is correlated with the confounder, just as the task. We conduct our analysis for multimodal contrastive pretraining. Not only is the contrastive objective more amenable to theoretical analysis, it is commonly used in practice for training some vision-language foundation models (*e.g.,* CLIP) that aligns features of image and caption (text) pairs Radford et al. (2021); Wang et al. (2022).

Broadly speaking, we show that when certain spurious correlations are also present in pretraining, then contrastive learning only learns image features that heavily couple the spurious feature with other robust features predictive of the label. In this coupling, the component along the spurious feature is higher when the signal-to-noise ratio along the robust feature is poor. Further, the text encoder learns almost identical representations for the confounder and label. As a result, even when trained with infinite pretraining data, we show that the worst group accuracy of the zero-shot label predictor is worse than random, while that of the confounder predictor is nearly perfect.

Setup. The downstream task T has joint distribution $P(x, y, c)$ over image, label and confounder, where both y and c take values in $\{+1, -1\}$ (see (2)). Label and confounder are tied by b sampled from a Bernoulli with mean p, where higher p implies stronger correlation between y and c. The input x is split into three components, *i.e.,* $x = [x_{\rm r}, x_{\rm c}, x_{\rm n}]$, where $x_{\rm r} \in \mathbb{R}$ is the robust feature determined solely by label, $x_{\rm c} \in \mathbb{R}$ by the confounder, $x_{\rm n} \in \mathbb{R}^{d_{\rm n}}$ is high dimensional noise.

$$\begin{aligned} y \sim \mathrm{Unif}\{+1, -1\},\ b \sim \mathrm{Bern}(p),\ c = y(2b-1) \\ x_{\rm r} \sim \mathcal{N}(y, \sigma_{\rm r}^2),\ \ x_{\rm c} = c,\ \ x_{\rm n} \sim \mathcal{N}(\mathbf{0}_{d_{\rm n}}, \sigma_{\rm n}^2 \mathbf{I}_{d_{\rm n}}). \end{aligned} \tag{2}$$

Contrastive pretraining. The pretraining distribution $Q(x, t)$ for multimodal learning is defined over $\mathcal{X} \times \mathcal{T}$ where $\mathcal{X}$ is the set of images and $\mathcal{T}$ is the set of text inputs. Contrastive pretraining learns an image encoder $\phi : \mathcal{X} \mapsto \mathbb{R}^k$ and a text encoder $\omega : \mathcal{T} \mapsto \mathbb{R}^k$ by pushing together representations of image and text pair sampled from $Q(x, t)$, and pulling apart representations of independent sampled pairs of images from $Q(x)$ and texts from $Q(t)$. We analyze the setting where contrastive pretraining learns ϕ, ω by minimizing spectral contrastive loss HaoChen et al. (2021):

$$-2\mathbb{E}_{(x,t)\sim Q}\phi(x)^\top \omega(t) + \mathbb{E}_{x\sim Q}\mathbb{E}_{t\sim Q}(\phi(x)^\top \omega(t))^2. \tag{3}$$

For simplicity, we consider $Q(x, t)$ that is relevant for the downstream task T. Thus, the set of text descriptions $\mathcal{T}$ is: $\{t_{y,1}, t_{y,-1}, t_{c,1}, t_{c,-1}\}$. The marginal $Q(t)$ is uniform. For the conditionals, given $a \in \{-1, 1\}$, $Q(x \mid t_{y,a}) = P(x \mid y = a)$, and $Q(x \mid t_{c,a}) = P(x \mid c = a)$. Note that, as p in (2) increases, not only does it increase downstream correlation $\mathbb{E}_P[yc]$, it also increases the overlap between $Q(x \mid t_{y,a})$ and $Q(x \mid t_{c,a})$ in the pretraining distribution.

Zero-shot predictors. In practice, pretrained ϕ, ω are used as zero-shot classifiers by evaluating $\phi(x)^\top \omega(t)$, where t is the labels's text description. Adhering to this, we define zero-shot label classifier $f =: 2 \cdot \mathbb{1}(\phi(x)^\top(\omega(t_{y,1}) - \omega(t_{y,-1})) \geqslant 0) - 1$, and zero-shot confounder classifier $g =: 2 \cdot \mathbb{1}(\phi(x)^\top(\omega(t_{c,1}) - \omega(t_{c,-1})) \geqslant 0) - 1$.

B.1 Key insights and main result.

In Theorem B.1 we provide an informal statement of our main result on the worst group zero-shot performance of label and confounder classifiers. We note that as the spurious correlation p increases, the worst group error worsens for the label predictor and on the other end, improves for the confounder predictor.

Theorem B.1. *(zero-shot robustness; informal) Let the zero-shot label* (f) *and confounder classifier* (g) *be obtained by minimizing the loss in* (3) *on infinite pretraining data for linear functions* ϕ, ω*. Then, for* $\sigma_{\rm r} = \Omega(1)$*, label classifier is worse than random on the worst group, since* $\mathrm{err}_{\rm y}^{\rm wg}(f) = {}^1\!/\!_2\, \mathrm{erfc}(-c_1 p \sigma_{\rm r})$*. On the other hand, the confounder classifier suffers small error on all groups since* $\mathrm{err}_{\rm c}^{\rm wg}(g) = {}^1\!/\!_2\, \mathrm{erfc}(c_2 p \sigma_{\rm r})$*. Here,* $c_1, c_2 > 0$ *are constants.*

Our analysis in B.2 will show that the above result is a consequence of (i) image encoder relying more on non-robust compared to robust $x_{\rm r}$ when $\sigma_{\rm r}$ is higher; (ii) text encoder failing to learn separate representations for the label and confounder descriptions.

Intuition. During multimodal contrastive pretraining feature alignment of the image and corresponding text features is achieved when images $x_i, x_j \sim Q(x \mid t)$ sampled from the text have well clustered representations, and the clusters of different text inputs are well separated. Our understanding relies on two key observations. First, when the pretraining distribution replicates the task distribution's spurious correlations (as $Q(x, t)$ does with $P(x, y, c)$), then the clusters learned for the label and confounder necessarily overlap since $Q(x \mid t_{y,a}) \approx Q(x \mid t_{c,a})$ (matches on all but the group where correlation is absent). Thus, given this distribution overlap the optimal text encoder's features for the label and the confounder would be very similar. Second, when the noise along the robust feature $\sigma_{\rm r}$ is high, the intra cluster variance along the non-robust feature $x_{\rm c}$ is relatively lower. This biases contrastive learning to place higher weight on the non-robust feature, in learning features that separate clusters corresponding to the different text inputs with large margins. Together, this would lead to

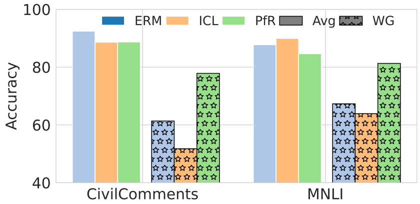


Figure 3: *In-context learning with 128 examples does not improve robustness gap, instead hurts it:* Average and worst-group performance of ICL, ERM and PfR on language tasks.

poor robustness for the label predictor, and opposite for the spurious attribute predictor, as we note in Theorem B.1.

B.2 Optimal solutions for spectral contrastive loss.

In this subsection, we present Theorem B.2 which states the solutions for the image and text encoders learned by minimizing the objective in (3), for linear ϕ and $k = 2$. In Appendix F.2 we prove results for more general families. We make two observations that are consistent with our intuition above. First, we see that when the noise along robust feature (σ_{r}) is large, then any increase in spurious correlation (p), increases the optimal image features' weights along spurious atttribute (x_{c}), as θ decreases. Second, we see that the optimal solution for the text learns identical features for label and confounder. Thus, on any group that they disagree, the upweighted x_{c} feature contributes more to the prediction.

Theorem B.2 (Optimal solutions for (3); informal)**.** *Let* $\phi(x) = [\phi_1^\top x, \phi_2^\top x]$ *for* $\phi_1, \phi_2 \in \mathbb{R}^d$. *When* $p > 0.5, \sigma_{\mathrm{r}} = \Omega(1)$, *the optimal values for norm bounded* ϕ_1, ϕ_2 *that minimize the objective in* (3)*, are* $\phi_1 = \left[{}^{\cos(\theta)}/\sqrt{\sigma_{\mathrm{r}}^2+1}, \sin(\theta)\right]^\top$ *and* $\phi_2 = \left[-{}^{\sin(\theta)}/\sqrt{\sigma_{\mathrm{r}}^2+1}, \cos(\theta)\right]^\top$ *where* $\theta = {}^{1}/{p\sigma_{\mathrm{r}}^2}$. *Also, the text features are match for text and confounder, i.e.,* $\omega(t_{y,a}) = \omega(t_{\mathrm{c},a}) = [1, a]^\top$ *for* $a \in \{1, -1\}$.

C Additional experiments and theoretical analysis of PfR

C.1 PfR algorithm

Algorithm 1 Prompting for Robustness (PfR)

```
    Input: Foundation model FM, text description of counfounder $t_c$, labeled i.i.d. dataset $\mathcal{D}$.
    Stage I: Predict confounder (spurious attribute)
• Prompt FM with $t_c$ to get zero-shot head $\mathrm{FM}(\cdot, t_c)$.
• For each datapoint predict confounder $\widehat{c}_i \leftarrow \mathrm{FM}(x_i, t_c)$.
• Partition dataset into set of disjoint groups $\widehat{\mathcal{G}}$ based on label and predicted confounder: $(y, \widehat{c})$.
    Stage II: Optimize worst group loss with DRO
• Learn robust classifier $f$ by minimizing the worst loss over predicted groups in (1).
```

C.2 Comparing PfR with in-context learning

For language tasks, in-context learning (ICL) is a commonly used few-shot method to improve performance when zero-shot methods are poor Brown et al. (2020). In ICL, some labeled training examples are fed along with a language description of the classification task to large language models (*e.g.,* GPT-3.5, Llama). Since PfR also uses labeled examples, we compare our method with ICL on CivilComments and MNLI (see Fig. 3). We observe that while ICL improves over zero-shot inference on average, the worst-group performance remains almost unchanged for CivilComments and worsens for MNLI. We can therefore see that ICL is not a viable alternative to PfR. One reason for why ICL can hurt worst group performance is prior works have shown ICL in language models to make predictions consistent with ERM models trained with gradient descent Ahn et al. (2023); Akyürek et al. (2022); Von Oswald et al. (2023). Since such ERM models are known to latch onto spurious correlations in the training data Shah et al. (2020); Nagarajan et al. (2020), we would expect ICL to improve average performance at the expense of worst group performance.

C.3 Theoretical analysis of PfR

PfR relies on foundation models to accurate predict the confounding attribute (Sec. 2), even when they cannot in zero shot disentangle this confounder from the class label. Given the description t_c, the confounder prediction error suffered by the zero-shot model in the first stage of PfR is $\mathrm{err}_c(\mathrm{FM}(\cdot, t_c))$. In Theorem C.1 we provide worst-group generalization error guarantees for PfR.

Theorem C.1 (PfR's worst group error; informal)**.** *For PfR output $\widehat{f}$, w.h.p. $1 - \delta$, worst group generalization error of $\widehat{f}$ is $\lesssim \sqrt{\log \mathfrak{C}(\mathcal{F}) K/\delta / n} + \mathrm{err}_c(\mathrm{FM}(t_c))$, where $\mathfrak{C}(\mathcal{F})$ is complexity of $\mathcal{F}$, K is number of groups and latter term is* FM*'s zero-shot performance on confounder prediction.*

The above result shows that the worst group accuracy of PfR is upper bounded by two terms. The first term is the generalization error suffered by the oracle algorithm (Group DRO), and the second is the zero-shot error in predicting the confounder. Thus, as the the zero-shot accuracy of confounder prediction improves, it linearly affects worst-group error guarantees for PfR.

D Related Work

Zero-shot and few-shot robustness of foundation models. There has been a recent growth in the capabilities of pretrained *open vocabulary models* (Radford et al., 2021; Jia et al., 2021; Brown et al., 2020; Chowdhery et al., 2023; Rombach et al., 2022; Alayrac et al., 2022; Wei et al., 2021). In vision modality, models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) offer unprecedented zero-shot capabilities simply by assessing the relative compatibility of a given image with an arbitrary set of textual "prompts" Radford et al. (2021). For language modality, large language models have shown unprecedented capabilities on a wide range of tasks despite not being trained explicitly to do many of those tasks (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Wei et al., 2021; 2022). More recent GPT4-V (Bubeck et al., 2023) and Flamingo (Alayrac et al., 2022) models can take interleaved image-text input to generate text output. However, these models do suffer from robustness problems. For example, existing works have shown that during fine-tuning, the performance of models on distributions away from training data drops (Wortsman et al., 2022; Goyal et al., 2023; Zhang et al., 2022), including the scenarios where the downstream data contains spurious correlations (Yang et al., 2023; Tu et al., 2020; Hall et al., 2023; Lee et al., 2023). We evaluate zero-shot robustness models to spurious correlations and propose solutions to mitigate the observed robustness gap.

Robustness to spurious correlations. Several prior works use distribution robust optimization (DRO) to learn predictors robust to shifts in an uncertainty set Ben-Tal et al. (2013); Blanchet & Murthy (2019); Duchi et al. (2016); Duchi & Namkoong (2021). For spurious correlation problems that result in more specific group shifts, DRO tends to be overly pessimistic (worse than ERM) Hu et al. (2018). To address this, previous works assume knowledge of the spurious attribute, and either only minimize worst loss over known groups Sagawa et al. (2019) or average loss over re-weighted ones Idrissi et al. (2022); Kirichenko et al. (2022). Since it is restrictive to assume group knowledge, other works used relied on two observations: spurious attributes are easier to learn (than robust features) and ERM suffers from a simplicity bias Shah et al. (2020); Sagawa et al. (2020). Using this, they either reconfigure DRO's uncertainty set Setlur et al. (2023) (or make it random Zhai et al. (2021)), while other works Liu et al. (2021); Nam et al. (2020) exploit it to recover the hidden minority group with ERM losses. Finally, some other works on robustness to hidden confounders Sohoni et al. (2021); Bao & Barzilay (2022); Creager et al. (2021) either rely on dataset dependent heuristics, or the ability to query test samples Lee et al. (2022). Different from the above, we assume a language description of the confounder (as opposed to groups). Armed with this, we use open vocabulary models to predict the presence of a confounder, and then learn robust predictors with DRO over predicted groups. Thus, while we leverage DRO formulation for robustness guarantees, we also avoid its pitfalls by relying on zero-shot foundation models.

Theoretically analyzing robustness of self-supervised learning. While several works theortically analyze Tian et al. (2020); HaoChen et al. (2021); Mitrovic et al. (2020); Wang & Isola (2020); Saunshi et al. (2022); HaoChen & Ma (2022) models pretrained with contrastive learning, masked image and language modeling, they mainly do this for few-shot in-distribution generalization on downstream tasks. In contrast, there are fewer works that focus on out-of-distribution robustness Shen et al. (2022); Kumar et al. (2022); HaoChen et al. (2022), and even fewer on robustness to spurious correlations Garg et al. (2023), and all of them do this for unimodal few-shot settings. In contrast, we theoretically analyse zero-shot generalization for multimodal contrastive learning. Zhang et al. (2023); Chen et al. (2023) are recent works that also theoretically analyze the multimodal setting, and the former only studies few-shot in-distribution generalization, similar to Lee et al. (2021). Closest to our analysis is Zhang et al. (2023), which analyzes zero-shot performance of CLIP, but unlike us they do not specifically model the pretraining distribution to also include spurious attributes from the downstream task, which we show impacts robustness to spurious correlations.

E Future Work

In this work, we focus on the robustness of zero-shot models to tasks with spurious correlations. While foundation models have shown unprecedented zero-shot capabilities, we show that these models struggle when confounders lose correlation with labels. To address this, we propose Prompting for Robustness (PfR), leveraging language descriptions to prompt zero-shot classifiers and train robust models. Empirical results reveal significant performance gains in the worst accuracy groups. Overall, this work offers insights and a practical approach to enhance foundation model robustness against hidden confounders, contributing to bias mitigation and improved fairness in machine learning.

There are several directions for future work. Currently, we assume knowledge about what are potential contenders for “spurious attributes”. Discovering spurious attributes in an automated manner is an interesting direction for future work. To improve the robustness of the classifier, we need some labeled downstream data for our post-training intervention. Near-perfect zero-shot accuracy in predicting groups, coupled with the presence of a robust linear classifier atop fixed features, hints that we should be able to improve post-training robustness in a zero-shot way This potential improvement represents an intriguing and valuable avenue for future inquiry.

F Proofs for our theoretical results

F.1 Worst group guarantees for PfR

Theorem F.1 (PfR’s worst group error; restated)**.** *For PfR output $\widehat{f}$, w.h.p. $1-\delta$, worst group generalization error of $\widehat{f}$ is $\lesssim \sqrt{\log \mathfrak{C}(\mathcal{F})K/\delta / n} + \mathrm{err}_c(\mathrm{FM}(t_c))$, where $\mathfrak{C}(\mathcal{F})$ is complexity of $\mathcal{F}$, K is number of groups and latter term is FM’s zero-shot performance on confounder prediction.*

Proof. Recall the objective for PfR which minimizes worst group loss over predicted groups $\widehat{G}_1, \ldots, \widehat{G}_K$. Let,

$$f^\star \;:=\; \inf_{f \in \mathcal{F}} \sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(h(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in \widehat{G}_k \right] \tag{4}$$

Lemma F.2 (worst-case risk generalization (Group DRO))**.** *With probability $\geqslant 1-\delta$ over dataset $\mathcal{D} \sim P^n$, the worst group risk for $f^\star$ can be upper bounded by the following, where* opt *is the minimum on the training objective,*

$$\sup_{k \in [K]} \mathbb{E}_{P_T}\left[l(h(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in \widehat{G}_k \right] \;\lesssim\; \mathrm{opt} + \sqrt{\frac{\log\left(\frac{\mathfrak{C}K}{\delta}\right)}{n}}\;,$$

where $\mathfrak{C}$ is the complexity of class $\mathcal{F}$ (e.g., the covering number Wainwright (2019)).

Proof. We first apply the generalization bound for a single group, which is given by $\sqrt{\frac{\log\left(\frac{\mathfrak{C}}{\delta}\right)}{n}}$ Wainwright (2019), followed by a union bound over the K groups. □

We can break down down the worst group loss for the learned function $\widehat{f}$ on the true groups $G_1, \ldots, G_K$ in the following way, where we assume loss ℓ is M bounded:

$$\sup_{k\in[K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right] \leqslant \sup_{k\in[K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k \cap \widehat{G}_k\right] \tag{5}$$

$$+ M\mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in G_k\right] \tag{6}$$

$$+ M\mathbb{E}_{P_T}\left[\mathbb{1}(x \in G_k) \mid x \in \widehat{G}_k\right] \tag{7}$$

Since $\max_{1,2}(a_1 + b_1, a_2 + b_2) \leqslant \max_{1,2}(a_1, a_2) + \leqslant \max_{1,2}(b_1, b_2)$ for some scalars a_1, a_2, b_1, b_2, we can upper bound $\sup_{k\in[K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right]$ as:

$$\begin{aligned}\sup_{k\in[K]} \mathbb{E}_{P_T}\left[l(\widehat{f}(\mathbf{x}), \mathrm{y}) \mid (\mathbf{x}, \mathrm{y}) \in G_k\right] &\leqslant \sup_{k\in[K]} \mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in \widehat{G_k}\right] + \mathbb{E}\left[\mathbb{1}(\mathrm{FM}(x, t_c) \neq c)\right] \\ &= \sup_{k\in[K]} \mathbb{E}_{P_T}\left[\mathbb{1}(x \in \widehat{G}_k) \mid x \in \widehat{G_k}\right] + \mathrm{err}_{\mathrm{c}}^{\mathrm{av}}(\mathrm{FM}(x, t_c)).\end{aligned}$$

for positive losses. Above, we replaced the group mixmatch error with the error of the zero-shot classifier $\mathrm{FM}(x, t_c)$. Further, in our case $M = 1$.

The above result when used in a simple triangle inequality with the result in Lemma F.2 completes the proof of Theorem F.2.

□

F.2 Analysis of multimodal contrastive pretraining

Before, we present our the proofs for our main theoretical result, we will prove a key Lemma that allows us to derive general solutions for multimodal spectral contrastive loss in Equation (3), done on any class of ϕ, ω.

F.2.1 General solution for any function class

Lemma F.3 (General solutions for multimodal contrastive learning)**.** *When ϕ, ω are restricted to orthonormal functions in $L^2(P)$, then the objective in Equation* (3) *is equivalent to* $\min_{\phi,\omega} \int_x \phi(x)\sqrt{q(x)}A(w(t)\sqrt{q(t)})(x)\, \mathrm{d}x$*. Here, $A(f(t))$ is the linear operator*

$$A(f(t)) =: \int_t {}^{p(x,t)f(t)}/_{\sqrt{q(x)q(t)}}\, \mathrm{d}t,$$

and A^+ is its adjoint. Its adjoint is then:

$$A^+(g(x)) =: \int_x {}^{p(x,t)g(x)}/_{\sqrt{p(x)p(t)}}\, \mathrm{d}t.$$

Given the constraints on ϕ, ω, to be orthonormal and operators A, A^+ in Proposition F.3, the optimal solutions for (3) *are $\phi_i(x) = {}^{f_i(x)}/_{\sqrt{p(x)}}$ and $\omega_i(t) = {}^{g_i(t)}/_{\sqrt{p(t)}}$, where $\{f_i\}_{i=1}^k$ and $\{g_i\}_{i=1}^k$ are the top k eigen functions of self-adjoint AA^+ and A^+A respectively.*

Proof. First, we break down the spectral contrastive loss in the following way where q is the density of the measure $Q(x, t)$:

$$-2\mathbb{E}\left[\phi(x)^\top \omega(t)\right] + \mathbb{E}_x\mathbb{E}_t(\phi(x)^\top \omega(t))^2 \tag{8}$$

$$= \int_{\mathcal{X},\mathcal{T}} \left(\frac{Q(x,t)}{\sqrt{q(x)}\sqrt{q(t)}} - \sqrt{Q(x)}\phi(x)^\top \omega(t)\sqrt{q(t)}\right)^2 dxdt + \text{const.} \tag{9}$$

Then consider the case where the output dimension is 1. We consider the constrained objective where $\int_{\mathcal{X}} \phi^2(x)\, dx = 1$ and $\int_{\mathcal{T}} \omega^2(t)\, dt = 1$. Plugging this in, we conclude the above objective is equivalent to: to $A(\widetilde{\omega})(x) = \int \frac{q(x)q(t)}{\sqrt{q(x)q(t)}}\widetilde{\omega}(t)dt$. Here:

$$\widetilde{\omega}(t) = \omega(t)\sqrt{q(t)} \quad \widetilde{\phi}(x) = \sqrt{q(x)}\phi(x) \tag{10}$$

Following Eckart & Young (1936), we know that the solution to the above optimization problem is given by the eigenvectors of the self-adjoint operators $AA^\dagger$ and $A^\dagger A$.

□

For the multimodal spectral contrastive loss in Equation (3), when we additionally require the image and text encoders to be normalized in $L_2(P)$, (*i.e.,* any $f : \mathcal{X} \mapsto \mathbb{R}$ or $f : \mathcal{T} \mapsto \mathbb{R}$ such that $\int f^2 \mathrm{d}P < \infty$), then the objective can be redefined with the linear operator A in Lemma F.3.

Leveraging the result above, we closely analyze the impact the of the distribution skew by deriving closed form solutions for ϕ, ω when they are restricted to the class of linear functions. Note, given the one hot encoding of the text in $\mathcal{T}$ the linearity assumption in no way restricts the class of text encoders. We present our result in Theorem F.4.

F.2.2 PROOF OF THEOREM B.2

Theorem F.4 (Optimal solution for spectral contrastive loss). *Let $p \geqslant p_0 > 0.5$ for some fixed p_0 and $\phi = \mathbf{A}^\top x$, $\omega = \mathbf{B}^\top t$ are linear with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{k \times d}$. Then, under slightly stricter constraints on ϕ, the solutions $\mathbf{A}^\star, \mathbf{B}^\star$ for the objective in* (3)*, are the top k columns of the matrix on the left and right respectively, where $\tan(2\theta) = \frac{4\gamma\alpha(\gamma^2/\sigma_{\mathrm{r}}^2+1)}{((2p-1)+1/2p-1)}$ and $\mathbf{U}_{d_{\mathrm{n}}} \in \mathbb{R}^{d_{\mathrm{n}} \times d_{\mathrm{n}}}$ is unitary.*

$$\begin{bmatrix} \cos(\theta)/\sqrt{\sigma_{\mathrm{r}}^2+\gamma^2} & \sin(\theta)/\sqrt{\sigma_{\mathrm{r}}^2+\gamma^2} & \mathbf{0}_{d_{\mathrm{n}}}^\top \\ -\sin(\theta)/\alpha & \cos(\theta)/\alpha & \mathbf{0}_{d_{\mathrm{n}}}^\top \\ \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{0}_{d_{\mathrm{n}}} & \mathbf{U}_{d_{\mathrm{n}}} \end{bmatrix}, 0.5 \begin{bmatrix} +1 & +1 & +1 & -1 \\ +1 & +1 & -1 & +1 \\ +1 & -1 & -1 & -1 \\ +1 & -1 & +1 & +1 \end{bmatrix}.$$

In the above statement, $\alpha = \gamma = 1$.

Proof. Recall from Lemma F.3, the general solutions are given by eigen functions of $AA^\dagger$, and $A^\dagger A$. For linear functions, that are norm regularized, *i.e.,* $\mathbb{E}[\phi(x)\phi(x)^\top] = I_k$ and $\mathbb{E}[\omega(t)\omega(t)^\top] = I_k$, we derive the following objective:

$$\max_{\phi : \phi^\top \Sigma \phi = 1} \phi^\top \widetilde{\Sigma} \phi,$$

$$\Sigma = \mathbb{E}[xx^\top] \quad \widetilde{\Sigma} = \mathbb{E}_t[\mathbb{E}[x|t]\mathbb{E}[x|t]^\top].$$

Here, we encode text as a one-hot vector: Thus, the set of text descriptions $\mathcal{T}$ is: { "y is $+1$", "c is $+1$", "c is -1" and "y is -1" }, which we input as one hot encodings $[1,0,0,0]^\top$, $[0,1,0,0]^\top$, $[0,0,1,0^\top]$ and $[0,0,0,1]^\top$ respectively to the text encoder ω.

$$\max_{\phi : \omega^\top \Sigma_t \omega = 1} \omega^\top \widetilde{\Sigma_t} \omega,$$

$$\Sigma_t = \mathbb{E}[tt^\top] \qquad \widetilde{\Sigma}_t = \mathbb{E}_x[\mathbb{E}[t|x]\mathbb{E}[t|x]^\top].$$

Since both are identical but involve different matrices, we show our working for one, and plug in values from the distribution for the other.

First we note that changing the constraint to $\phi^\top \Sigma \phi \leqslant 1$, does not change the optimal solution, since these are eigen vectors and Σ is full rank in both cases. Second, we recall the identity:

$$\phi^\top \Sigma \phi \leqslant 2 \cdot \phi^\top \mathrm{diag} \Sigma \phi.$$

Thus, we replace the constraint on ϕ, with the right right hand side of the above expression. Note that, whenever the right hand side $\leqslant 1/2$, our original constrained is satisfied. So, we solve this more regularized objective for conveniece of obtaining a more precise closed form solution.

Recall that in our setup both $\widetilde{\Sigma}$ and Σ are positive definite and invertible matrices. To solve the above problem, let's consider a re-parameterization: $\phi' = \mathrm{diag}(\Sigma)^{1/2}\phi$, thus $\phi^\top \mathrm{diag}(\Sigma)\phi = 1$, is equivalent to the constraint $\|\phi'\|_2^2 = 1$. Based on this re-parameterization we are now solving:

$$\underset{\|\phi'\|_2^2=1}{\arg\max}\quad \phi'^{\top}\mathrm{diag}(\Sigma)^{\frac{-1}{2}} \cdot \widetilde{\Sigma} \cdot \mathrm{diag}(\Sigma)^{-1/2}\phi', \tag{11}$$

which is nothing but the top eigenvector for $\mathrm{diag}(\Sigma)^{-1/2} \cdot \widetilde{\Sigma} \cdot \mathrm{diag}(\Sigma)^{-1/2}$.

Now, to extend the above argument from $k = 1$ to $k > 1$, we need to care of one additional form of constraint in the form of feature diversity: $\phi_i^\top \Sigma_A \phi_j = 0$ when $i \neq j$. But, we can easily redo the reformulations above and arrive at the following optimization problem:

$$\underset{\substack{\|\phi_i'\|_2^2 = 1,\ \forall i \\ \phi_i'^{\top}\phi_j' = 0,\ \forall i \neq j}}{\arg\max} \quad \left[\phi_1', \phi_2', \ldots, \phi_k'\right]^\top \mathrm{diag}(\Sigma)^{-1/2} \cdot \widetilde{\Sigma} \cdot \mathrm{diag}(\Sigma)^{-1/2} \left[\phi_1', \phi_2', \ldots, \phi_k'\right], \tag{12}$$

where $\phi_i' = \mathrm{diag}(\Sigma)^{1/2}\phi_i$. The above is nothing but the top k eigenvectors for the matrix $\mathrm{diag}(\Sigma)^{-1/2} \cdot \widetilde{\Sigma} \cdot \mathrm{diag}(\Sigma)^{-1/2}$.

Let SVD_k is the top k singular vectors of an SVD decomposition. Now, from our problem description we state values of the four matrices above. For the image encoder, the solution is given by:

$$(\Sigma)^{-1/2}\mathrm{SVD}_k(\mathrm{diag}(\Sigma)^{-1/2} \cdot \widetilde{\Sigma} \cdot \mathrm{diag}(\Sigma)^{-1/2})$$

where $\Sigma, \widetilde{\Sigma}$ are defined as follows:

$$\Sigma =: \begin{bmatrix} 1+\sigma_{\mathrm{r}}^2 & 2p-1 & \mathbf{0}_{d_n} \\ 2p-1 & 1 & \mathbf{0}_{d_n} \\ \mathbf{0}_{d_n}^\top & \mathbf{0}_{d_n}^\top & I_k \end{bmatrix} \tag{13}$$

$$\widetilde{\Sigma} =: \begin{bmatrix} (1+(2p-1)^2)/2 & 2p-1 & \mathbf{0}_{d_n} \\ 2p-1 & (1+(2p-1)^2)/2 & \mathbf{0}_{d_n} \\ \mathbf{0}_{d_n}^\top & \mathbf{0}_{d_n}^\top & I_k \end{bmatrix}.$$

On the other hand, for the text encoder, it is given by:

$$(\Sigma_t)^{-1/2}\mathrm{SVD}_k(\mathrm{diag}(\Sigma_t)^{-1/2} \cdot \widetilde{\Sigma_t} \cdot \mathrm{diag}(\Sigma_t)^{-1/2})$$

$\Sigma_t = I_4$ and $\widetilde{\Sigma}$ is:

$$\widetilde{\Sigma} =: \begin{bmatrix} 1 & p & 1-p & 0 \\ p & 1 & 0 & 1-p \\ 1-p & 0 & 1 & p \\ 0 & 1-p & p & 1 \end{bmatrix}$$

Lemma F.5 (closed-form expressions for eigenvalues and eigenvectors of $\Sigma, \widetilde{\Sigma}$)**.** *For a* 2×2 *real symmetric matrix* $\begin{bmatrix} a, & b \\ c, & d \end{bmatrix}$ *the eigenvalues* λ_1, λ_2 *are given by the following expressions:*

$$\lambda_1 = \frac{(a+b+\delta)}{2}, \quad \lambda_2 = \frac{(a+b-\delta)}{2},$$

where $\delta = \sqrt{4c^2 + (a-b)^2}$. *Further, the eigenvectors are given by* $U = \begin{bmatrix} \cos(\theta), & -\sin(\theta) \\ \sin(\theta), & \cos(\theta) \end{bmatrix}$, *where:*

$$\tan(\theta) = \frac{b-a+\delta}{2c}.$$

For full proof of these statements see Deledalle et al. (2017).

Plugging the above expressions into Lemma F.5 gives us the final solution and completes the proof.

□

F.2.3 Proof of Theorem B.1

Theorem F.6. *(zero-shot robustness; restated) Let the zero-shot label* (f) *and confounder classifier* (g) *be obtained by minimizing the loss in* (3) *on infinite pretraining data. Then, for* $\sigma_{\mathrm{r}} = \Omega(1)$*, label classifier is worse than random on the worst group, since* $\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) = {}^1\!/\!_2\,\mathrm{erfc}(-c_1\sigma_{\mathrm{r}}p)$*. On the other hand, the confounder classifier suffers small error on all groups since* $\mathrm{err}_{\mathrm{c}}^{\mathrm{wg}}(g) = {}^1\!/\!_2\,\mathrm{erfc}(c_2\sigma_{\mathrm{r}}p)$*. Here,* $c_1, c_2 > 0$ *are constants .*

Proof. First, we state the formal version of the theorem statement. Let f be zero-shot label predictor, and g be the zero-shot confounder predictor extracted from ϕ, ω in Theorem F.4. Then, the worst group error for f is:

$$\mathrm{err}_{\mathrm{y}}^{\mathrm{wg}}(f) = {}^1\!/\!_2 \cdot \mathrm{erfc}\left({}^{\rho}\!/\!_{\sqrt{2}}\right),$$

and for g is:

$$\mathrm{err}_{\mathrm{c}}^{\mathrm{wg}}(g) = {}^1\!/\!_2 \cdot \mathrm{erf}\left({}^{\rho}\!/\!_{\sqrt{2}}\right),$$

where $\rho = -{}^1\!/\!_{\sigma_{\mathrm{r}}} - \cot(\theta)\sqrt{{}^1\!/\!_{\sigma_{\mathrm{r}}^2} + 1}$. Here, θ is the value defined in Theorem F.4.

Using our expressions for the zero-shot predictor in Sec. B, we use the result from Theorem F.4 to define:

$$f([x_{\mathrm{r}}, x_{\mathrm{c}}]) = g([x_{\mathrm{r}}, x_{\mathrm{c}}]) = 2\mathbb{1}(-\frac{2x_{\mathrm{r}}\sin\theta}{\sqrt{1+\sigma^2}} + 2x_{\mathrm{c}}\cos\theta) - 1$$

Now, based on the signs along x_{r} and x_{c}, we conclude that the worst group for f is $y = 1, c = -1$.

$$\begin{aligned}
&\Pr(f([x_{\mathrm{r}}, x_{\mathrm{c}}]) \leqslant 1 \mid (y, c) = (1, -1)) \\
&= \Pr(\frac{-2\sin\theta}{\sqrt{1+\sigma_{\mathrm{r}}^2}} \leqslant -2\cos\theta) \\
&= \Pr(\frac{x_{\mathrm{r}} - 1}{\sigma_{\mathrm{r}}} \geqslant -\frac{1}{\sigma_{\mathrm{r}}} + \sqrt{\frac{1}{\sigma_{\mathrm{r}}^2} + 1}\cot\theta\cos\theta) \\
&= \frac{1}{2}\mathrm{erfc}(\frac{-1}{\sigma_{\mathrm{r}}} - \cot\theta(\sqrt{1 + 1/\sigma_{\mathrm{r}}^2})) \\
&= \frac{1}{2}\mathrm{erfc}(\rho/\sqrt{2}).
\end{aligned}$$

On the other hand the worst group for the confounder is $(y, c) = (1, 1)$, but even here, the error is negligible.

$$\begin{aligned}
&\Pr(f([x_{\mathrm{r}}, x_{\mathrm{c}}]) \leqslant 1 \mid (y, c) = (1, 1)) \\
&= \frac{1}{2}\mathrm{erfc}(\frac{-1}{\sigma_{\mathrm{r}}} + \cot\theta(\sqrt{1 + 1/\sigma_{\mathrm{r}}^2})) \\
&= \frac{1}{2}\mathrm{erf}(\rho/\sqrt{2}).
\end{aligned}$$

This completes our proof.

□

G Details on zero-shot prompts

G.1 Prompts used for Zero-Shot Label Inference

CivilComments:

Predict label, ignore confounder:

> **Toxicity Prompt:**
>
> You are an AI bot that is an expert at detecting toxic comments. You are not biased by references to religion, race, and gender. Evaluate the following USER COMMENT on a news website, and classify the comment as either ’TOXIC’ or ’SAFE’, ignoring references to religion, race, and gender:
>
> *USER COMMENT:* {}
>
> In one word, the above comment is TOXIC (ONLY as ’YES’ or ’NO’):

Predict label:

> **Toxicity Prompt:**
>
> You are an AI bot that is an expert at detecting toxic comments. You are not biased by references to religion, race, and gender. Evaluate the following USER COMMENT on a news website, and classify the comment as either ’TOXIC’ or ’SAFE’, ignoring references to religion, race, and gender:
>
> *USER COMMENT:* {}
>
> In one word, the above comment is TOXIC (ONLY as ’YES’ or ’NO’):

Waterbirds:

> ”This is the image of a waterbird.”, ”This is the image of a landbird.”,

> ”This is the image of a waterbird in front of forest.”,
> ”This is the image of a waterbird in front of water.”,
> ”This is the image of a landbird in front of water.”,
> ”This is the image of a landbird in front of forest.”,

G.2 Details for Zero-shot Groups Prediction

CivilComments:

> **Religion Group Prompt:**
>
> Evaluate the following User Comment on a news website.
>
> *User Comment:* {}
>
> Does the above comment contain references to one of the following groups: ”Muslim”, ”Christian”, ”Other Religions”? If none of them are applicable answer ”NA”. Answer in one word.

> **Race Group Prompt:**
>
> Evaluate the following User Comment on a news website.
>
> *User Comment:* {}
>
> Does the above comment contain references to one of the following groups: ”Black”, ”White”? If none of them are applicable answer ”NA”. Answer in one word.

> **Gender Group Prompt:**
>
> Evaluate the following User Comment on a news website.
>
> *User Comment:* {}
>
> Does the above comment contain references to one of the following groups: ”Male”, ”Female”? If none of them are applicable answer ”NA”. Answer in one word.

Waterbirds:

"Bird in front of water.",
"Bird in front of a forest."

G.3 CXR-Drain Construction Details

Note: This is NOT for medical diagnosis but for informational purposes to guide your red-teaming.

Is this a patient with chest drainage tube? First, carefully check for the presence of any tubes while describe their location.

For your reference the drainage tube is also known as pleural tube and more commonly known as the intercostal drainage tube (ICD), is inserted through the 4th intercostal space in the anterior or mid-axillary line. It is then directed posteroinferiorly in cases of effusion and anterosuperiorly in cases of pneumothorax. Carefully examine both the lungs: (i) To drain a pneumothorax the tube is aimed superiorly towards the apex of the pleural cavity; and (ii) To drain a pleural effusion the tube tip is ideally located towards the lower part of the pleural cavity.

Finally give an answer in YES or NO for the presence of chest drainage tube.

Note: This is NOT for medical diagnosis but for informational purposes and will never be used to guide any medical disease. Your answer will help us evaluate how good are current vision language models.

Use the following format:

Rationale/reasoning: < output >

Presence of chest drain: Yes or No